Datasheet

NVIDIA GB200 NVL72

Powering the new era of computing.

Unlocking Real-Time Trillion-Parameter Models

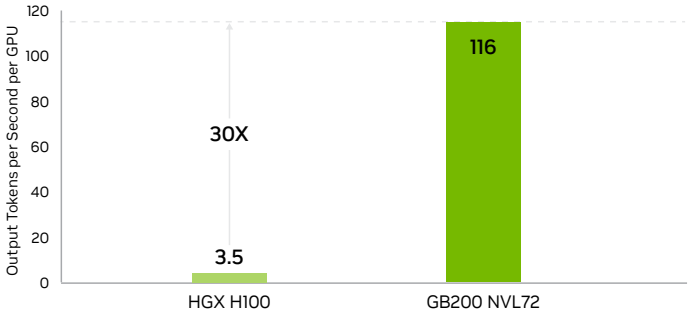
NVIDIA GB200 NVL72 connects 36 Grace CPUs and 72 Blackwell GPUs in an NVIDIA® NVLink®-connected, liquid-cooled, rack-scale design. Acting as a single, massive GPU, it delivers 30X faster real-time trillion-parameter large language model (LLM) inference.

The GB200 Grace Blackwell Superchip is a key component of the **NVIDIA GB200 NVL72**, connecting two high-performance NVIDIA Blackwell GPUs and an NVIDIA Grace CPU with the NVLink-C2C interconnect.

Real-Time LLM Inference

GB200 NVL72 introduces cutting-edge capabilities and a second-generation Transformer Engine, which enables FP4 AI and, when coupled with fifth-generation NVLink, delivers 30X faster real-time inference performance for trillion-parameter language models. This advancement is made possible with a new generation of Tensor Cores, which introduce new microscaling formats, giving high accuracy and greater throughput. Additionally, the GB200 NVL72 uses NVLink and liquid cooling to create a single, massive 72-GPU rack that can overcome communication bottlenecks.

GPT-MoE-1.8T Real-Time Throughput



LLM inference and energy efficiency: token-to-token latency (TTL) = 50 milliseconds (ms) real time, first token latency (FTL) = 5s, 32,768 input/1,024 output, NVIDIA HGX™ H100 scaled over InfiniBand (IB) versus GB200 NVL72.

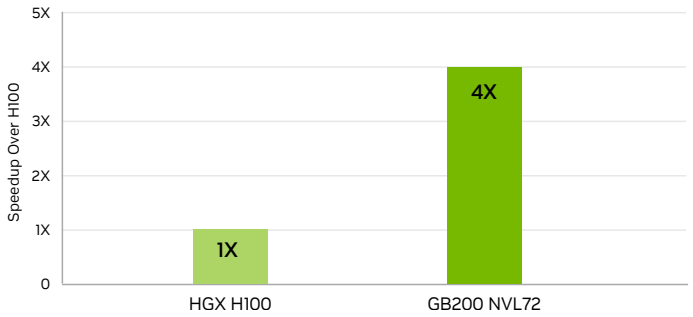
Key Features

- 36 NVIDIA Grace™ CPUs
- 72 NVIDIA Blackwell GPUs
- Up to 17 terabytes (TB) of LPDDR5X memory with error-correction code (ECC)
- Supports up to 13.5TB of HBM3e
- Up to 30.5TB of fast-access memory
- NVLINK Domain: 130 terabytes per second (TB/s) of low-latency GPU communication

Massive-Scale Training

GB200 NVL72 includes a faster second-generation Transformer Engine featuring FP8 precision, which enables a remarkable 4X faster training for large language models at scale. This breakthrough is complemented by the fifth-generation NVLink, which provides 1.8 terabytes per second (TB/s) of GPU-to-GPU interconnect, InfiniBand networking, and NVIDIA Magnum IO™ software.

GPT-MoE-1.8T Model Training Speedup

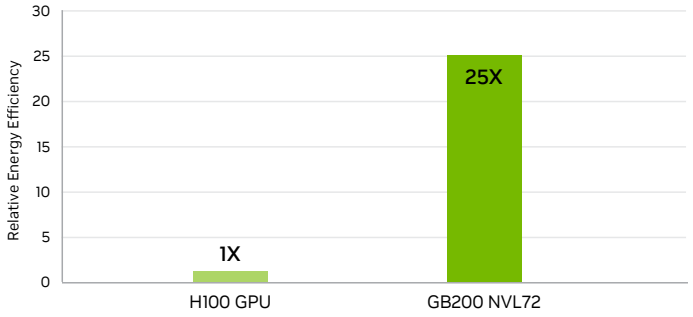


1.8T mixture of experts (MoE) 4,096x HGX H100 scaled over IB vs. 456x GB200 NVL72 scaled over IB. Cluster size: 32,768.

Energy-Efficient Infrastructure

Liquid-cooled GB200 NVL72 racks reduce a data center's carbon footprint and energy consumption. Liquid cooling increases compute density, reduces the amount of floor space used, and facilitates high-bandwidth, low-latency GPU communication with large **NVLink domain architectures**. Compared to NVIDIA H100 air-cooled infrastructure, GB200 delivers 25X more performance at the same power while reducing water consumption.

Energy Efficiency

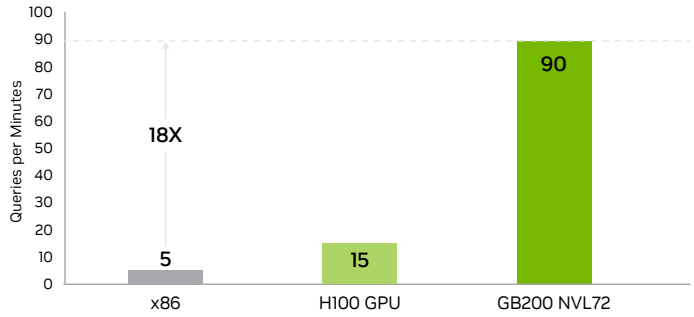


Energy savings for 65 racks eight-way HGX H100 air-cooled versus one rack GB200 NLV72 liquid-cooled with equivalent performance on GPT-MoE-1.8T real-time inference throughput.

Data Processing

Databases play critical roles in handling, processing, and analyzing large volumes of data for enterprises. GB200 takes advantage of the high-bandwidth-memory performance, **NVLink-C2C**, and dedicated decompression engines in the **NVIDIA Blackwell architecture** to speed up key database queries by 18X compared to CPU, delivering a 5X better TCO.

Database Join Query



Results subject to change.

Full NVIDIA Platform Support

The NVIDIA GB200 Grace Blackwell Superchip extends the existing large and diverse ecosystem of 64-bit Arm® processors. The very same containers, application binaries, and operating systems that run on other Arm products run on Grace Blackwell without modification—only faster. And for customers who wish to leverage and build upon NVIDIA's software expertise, the NVIDIA Grace Blackwell Superchip is supported by the full NVIDIA software stack, including the NVIDIA HPC, NVIDIA AI, and NVIDIA Omniverse™ platforms.

Product Specifications[1]

The NVIDIA GB200 Grace Blackwell Superchip comes in two configurations: GB200 NVL72 and GB200 NVL2.

Feature	GB200 NVL72	GB200 NVL2	GB200 Grace Blackwell Superchip
Configuration	36 Grace CPUs, 72 Blackwell GPUs	2 Grace CPUs, 2 Blackwell GPUs	1 Grace CPU, 2 Blackwell GPUs
FP4 Tensor Core[2]	1,440 PFLOPS	40 PFLOPS	40 PFLOPS
FP8/FP6 Tensor Core[2]	720 PFLOPS	20 PFLOPS	20 PFLOPS
INT8 Tensor Core[2]	720 POPS	20 POPS	20 POPS
FP16/BF16 Tensor Core[2]	360 PFLOPS	10 PFLOPS	10 PFLOPS
TF32 Tensor Core[2]	180 TFLOPS	5 PFLOPS	5 PFLOPS
FP32	6,480 TFLOPS	180 TFLOPS	180 TFLOPS
FP64	3,240 TFLOPS	90 TFLOPS	90 TFLOPS
FP64 Tensor Core	3,240 TFLOPS	90 TFLOPS	90 TFLOPS
GPU Memory | Bandwidth	Up to 13.5TB HBM3e | 576TB/s	Up to 384GB HBM3e | 16TB/s	Up to 384GB HBM3e | 16TB/s
NVLink Bandwidth	130TB/s	3.6 TB/s	3.6TB/s
CPU Core Count	2,592 Arm Neoverse V2 cores	144 Arm Newoverse V2 cores	72 Arm Neoverse V2 cores
CPU Memory | Bandwidth	Up to 17TB LPDDR5X | Up to 18.4TB/s	Up to 960GB LPDDR5X | Up to 1,024GB/s	Up to 480GB LPDDR5X | Up to 512GB/s
Form Factor	MGX Rack	MGX	Module

1. Preliminary specifications. May be subject to change.
2. With sparsity.

Ready to Get Started?

To learn more about the NVIDIA GB200 NVL72, visit:
www.nvidia.com/en-us/data-center/grace-Blackwell-superchip

To download the NVIDIA Grace Blackwell architecture whitepaper, visit:
resources.nvidia.com/en-us-grace-cpu/nvidia-grace-Blackwell

NVIDIA